Data Privacy

Garantia de la Informació i Seguretat

Departament d'Enginyeria de la Informació i les Comunicacions
Universitat Autònoma de Barcelona

1

Privacy (and anonymity)

- A very wide concept:
  - Privacy in communication networks
  - Data privacy (document confidentiality, private data publishing…)
  - Lots of other more
- **Privacy Enhancing Technologies (PET)**
  - used to widely refer to technologies used to provide privacy.

2

Some examples of PETs

- The TOR network,
- Mix networks (Mixnets),
- Encryption,
- Use of pseudonyms,
- Anonymous credentials,
- Location privacy,
- …

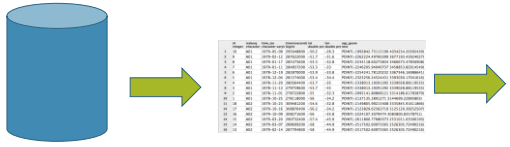
3

Private information publishing

- We will deal with a specific privacy problem:
  data publishing
- Includes several disciplines, but mainly:
  - Statistical Disclosure Control
  - Privacy Preserving Data mining

4

Motivating example



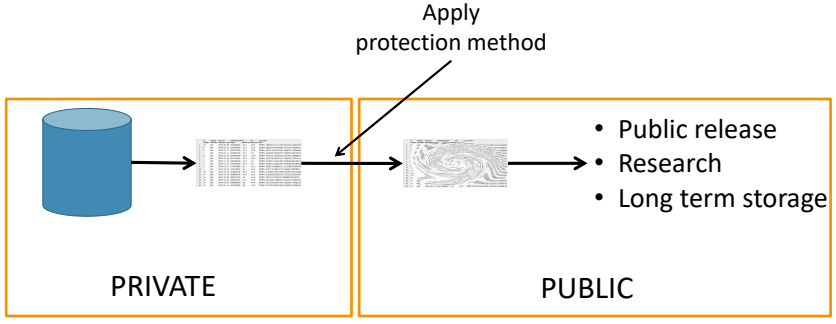
Database with **private** data or information

Data publication

- Public release
- Research (marketing, science, …)
- Long term storage

5

Need to ensure privacy



6

Definitions: anonymity

Anonymity of a subject means that the subject is not identifiable with a set of subjects, the **anonymity set**.

Unlinkability of two or more items of interest (e.g. subjects, messages, actions, …) from an attacker' s perspective means that within the system, the attacker cannot sufficiently distinguish whether these items are related or not

7

Definitions: disclosure

Disclosure takes place when attackers take advantages of the observation of available data to improve their knowledge on some confidential information about an item of interest.

- **Identity disclosure**
  - The attacker can correctly identify a particular entity in the system, or can link some protected data to an specific entity.
- **Attribute disclosure**
  - The attacker can learn something new about an attribute of an entity.

8

Example of disclosure

- This table has been naively anonymized by removing identifier attributes (name, national ID number, etc.)

City	Age	Profession	Income
Campeche	30	Teacher	200
Campeche	30	Teacher	250
Campeche	30	Teacher	300
Palenque	27	Physician	350
Palenque	45	Police	250

- But, what happens if an attacker knows:
  1. Alice is 30 years old and lives in Campeche
  2. Bob is 45 years old and lives in Palenque

9

Example 2

Name	Passport	Profession	City	Age	Salary
Alfredo Pareja	AS3682	Teacher	Guayaquil	23	600
Eugenio Espejo	RX3453	Teacher	Quito	48	500
Raul Perez	TS5645	Policeman	Loja	20	734
Francisco Proaño	XB3456	Writer	Guayaquil	32	543
Cesar Davila	AE4324	Teacher	Cuenca	38	890
Ignacio Laso	FF3455	Writer	Quito	41	678
Raul Serrano	LO0903	Policeman	Portoviejo	21	399
Luz Argentina	KO8264	Policeman	Quevedo	44	943
Abdon Ubidia	RC3766	Teacher	Machala	31	632

10

Example 2: naïve anonymization

Are you from Quevedo? Identity disclosure

Are you from Guayaquil? Attribute disclosure

Name	Passport	Profession	City	Age	Salary
XXXXXX	XXXXXX	Teacher	Guayaquil	32	600
XXXXXX	XXXXXX	Teacher	Quito	48	500
XXXXXX	XXXXXX	Teacher	Loja	20	734
XXXXXX	XXXXXX	Writer	Guayaquil	32	543
XXXXXX	XXXXXX	Teacher	Cuenca	38	890
XXXXXX	XXXXXX	Writer	Quito	41	678
XXXXXX	XXXXXX	Policeman	Portoviejo	21	399
XXXXXX	XXXXXX	Policeman	Quevedo	44	943
XXXXXX	XXXXXX	Teacher	Machala	31	632

11

Example: AOL Case

- AOL publicly releases in 2006: query logs by 650,000 users over 3 months
- Logs where apparently anonymized

```
24969 orioles tickets 2006-05-31 12:31:57 2 http://www.greatseats.com
24969 jennifer craford my space.com 2006-05-31 19:15:02
24969 jennifer crawford my space.com 2006-05-31 19:16:05
14423 boston redsoxweb page.com 2006-03-28 17:51:55
14423 www.bostonredsox 2006-03-28 18:12:26 1 http://boston.redsox.mlb.com
...
```


12

Example: AOL case

- A user (Thelma Arnold) was easily **re-identified**

- https://www.nytimes.com/2006/08/09/technology/09aol.html
- https://en.wikipedia.org/wiki/AOL_search_data_leak

HOME PAGE | MY TIMES | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times — Technology

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

CAMCORDERS | CAMERAS | CELLPHONES | COMPUTERS | HANDHELDS | HOME VIDEO | MUSIC | PERIPH

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

SIGN IN TO E-MAIL THIS | PRINT | SINGLE PAGE | REPRINTS

 Erik S. Lesser for The New York Times
Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

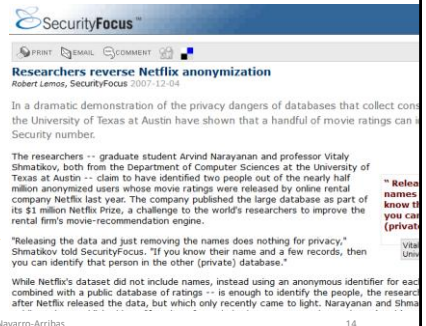
And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

13

Example: Netflix

- Netflix:
  - $1M prize for 10% improvement of its recommendation system.
  - Released an *anonymized* data to be used as training set.
- Researches could re-identify users by linking the anonymized training data to IMDB reviews!
- https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf



SecurityFocus

PRINT | EMAIL | COMMENT

Researchers reverse Netflix anonymization

Robert Lemos, SecurityFocus 2007-12-04

In a dramatic demonstration of the privacy dangers of databases that collect cons... the University of Texas at Austin have shown that a handful of movie ratings can ... Security number.

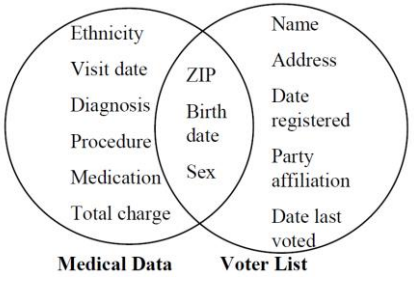
The researchers -- graduate student Arvind Narayanan and professor Vitaly Shmatikov, both from the Department of Computer Sciences at the University of Texas at Austin -- claim to have identified two people out of the nearly half million anonymized users whose movie ratings were released by online rental company Netflix last year. The company published the large database as part of its $1 million Netflix Prize, a challenge to the world's researchers to improve the rental firm's movie-recommendation engine.

"Releasing the data and just removing the names does nothing for privacy," Shmatikov told SecurityFocus. "If you know their name and a few records, then you can identify that person in the other (private) database."

While Netflix's dataset did not include names, instead using an anonymous identifier for eac... combined with a public database of ratings -- is enough to identify the people, the researc... after Netflix released the data, but which only recently came to light. Narayanan and Shma...

14

Example: US Census and Medical data



Researcher found that:
- date of birth
- gender
- 5-digit ZIP

Uniquely identifies 87.1% of USA population!!!

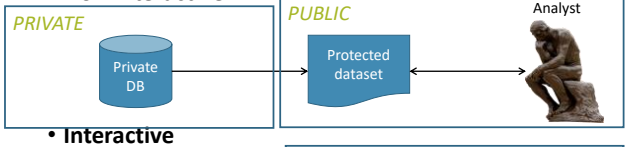
https://dataprivacylab.org/projects/identifiability/paper1.pdf

- Link two public anonymized datasets

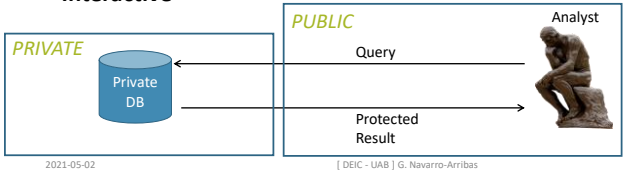
15

Privacy Protection Scenarios

- **Non-interactive**



- **Interactive**



Analysis:
- Statistics
- Data mining
- Machine learning
- etc.

16

Privacy Models

- k-anonymity
- Differential Privacy

17

Type of attributes related to an entity

- **Identifiers**: unambiguously identify the entity (passport number, full name, etc.)
  - Usually removed or encrypted for public release.
- **Quasi-identifiers**: in combination, can be linked with external information to reidentify an entity (age, zip code, city, …)
  - Usually a protection mechanism is applied to these attributes.
- **Confidential**: contain sensitive information about the entity (salary, religion, political affiliation, health condition, etc.)
  - Usually not modified (object of study).

18

Example

Identifiers | Quasi-identifiers | Confidential

Name	Passport	Profession	City	Age	Salary
Alfredo Pareja	AS3682	Teacher	Guayaquil	23	600
Eugenio Espejo	RX3453	Teacher	Quito	48	500
Raul Perez	TS5645	Policeman	Loja	20	734
Francisco Proaño	XB3456	Writer	Guayaquil	32	543
Cesar Davila	AE4324	Teacher	Cuenca	38	890
Ignacio Laso	FF3455	Writer	Quito	41	678
Raul Serrano	LO0903	Policeman	Portoviejo	21	399
Luz Argentina	KO8264	Policeman	Quevedo	44	943
Abdon Ubidia	RC3766	Teacher	Machala	31	632

19

Definition of k-anonymity

A dataset X satisfies k-anonymity with respect to a set of quasi-identifiers when the projection of X in this set results into a partition of X in sets of at least k indistinguishable records

- There are **k** indistinguishable items
- An item (individual) cannot be distinguished from other k-1 items from the dataset.

20

Example of k-anonymity

Quasi-identifiers | Confidential

City	Age	Salary
Guayaquil	23	600
Quito	48	500
Loja	20	734
Guayaquil	32	543
Cuenca	38	890
Quito	41	678
Portoviejo	21	399
Quevedo	44	943
Machala	31	632

3-anonymous table →

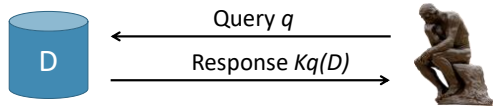
City	Age	Salary
(Ecuador)	[20, 29]	600
(Ecuador)	[40, 49]	500
(Ecuador)	[20, 29]	734
(Ecuador)	[30, 39]	543
(Ecuador)	[30, 39]	890
(Ecuador)	[40, 49]	678
(Ecuador)	[20, 29]	399
(Ecuador)	[40, 49]	943
(Ecuador)	[30, 39]	632

21

Differential Privacy

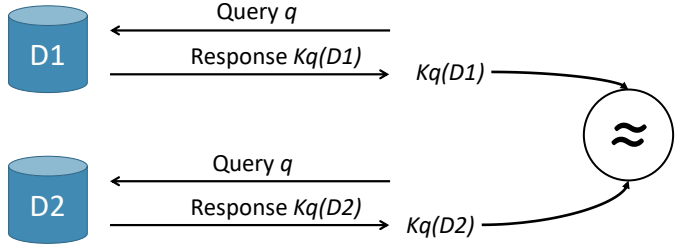
A function K_q for a query q gives ε-differential privacy if for all data sets D_1 and D_2 differing in at most one element, and all $S \subseteq Range(K_q)$,

$$\frac{\Pr[K_q(D_1) \in S]}{\Pr[K_q(D_2) \in S]} \leq e^{\varepsilon}$$



22

(Very) broadly speaking…



23

Differential privacy

- For any D_1, D_2: two versions of the database *D* that differ in at most one item (element, record, …)
- If we perform the same query to both, the result should be approximately the same.
- ε is the level of privacy (smaller ε ⟹ greater privacy). E.g. for values:
  - 0: total privacy (both queries return the same answer)
  - 0.01: both probabilities differ by <1%
  - 1: both probabilities differ by *e*

24

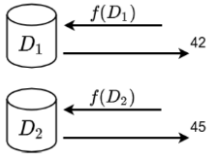
Differential privacy example

Table 1: D_1

nombre	edad	salario
Ataúlfo	43	100
Sigerico	15	300
Walia	50	200
Teodoredo	51	100
Turismundo	35	500
Tedorico II	40	300
Eurico	44	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	24	200

Table 2: D_2

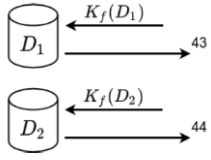
nombre	edad	salario
Ataúlfo	43	100
Walia	50	200
Teodoredo	51	100
Turismundo	35	500
Tedorico II	40	300
Eurico	44	700
Alarico II	49	300
Gesaleico	69	500
Amalarico	24	200



25

Cont.

- Imagine we have *Kf = f + noise*. Noise is a number randomly chosen from *{-2, -1, 0, 1, 2}*.



26

Privacy budget in differential privacy

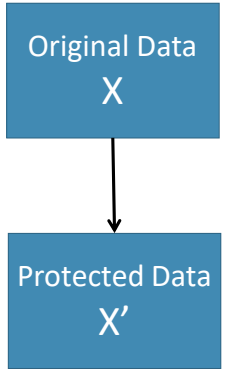
- In differential privacy the loss of privacy is accumulative
  - E.g. the same query to the same DB:
  - One query: 1/5 probability of guessing the real value.

Consulta	$K_f(D_1)$	¿ $f(D_1)$?
1	43	$41, 42, 43, 44, 45$
2	40	$38, 39, 40, 41, 42$

  - Now, we know the real values is 41 or 42, prob: 1/2
- We have a privacy budget which limits the number of queries!!

27

Protection methods



28

Protection methods classification

- **Perturbative**: the original data X is distorted. Protected data X' contain some erroneous information.
- **Non-perturbative**: replace original values X by others which are less specific. Protected data X' do not contain erroneous information.
- **Synthetic data generators**: new artificial data is generated to substitute the original data.

Some protection methods

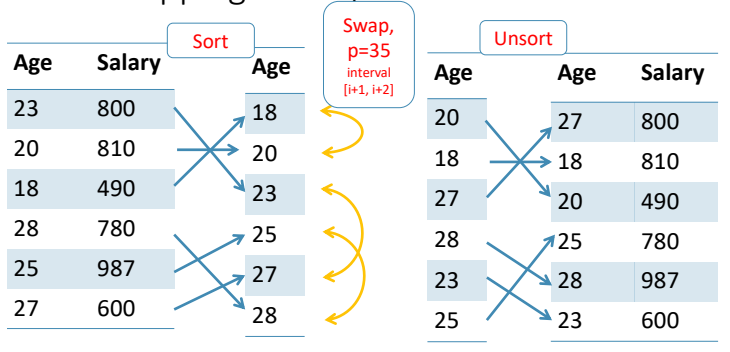
We will see some examples of common protection methods

- Perturbative
  - Rank Swapping
  - Microaggregation
  - Additive and multiplicative Noise
- Non-perturbative
  - Generalization

Rank Swapping

- For each single variable V:
  1. Sort values $V = (a_1, \dots, a_n)$, such that $a_i \leq a_j$ for all $1 \leq i \leq j \leq n$
  2. Swap each value a_i with a_l, where l is randomly chosen from $[i + 1, \min(n, i + p * |X|/100)]$
  3. Undo the sorting step (1)
- Each value is swapped with another one randomly chosen within a restricted range, determined by *p*.
- *p* is the percentage from the total number of values.

Rank swapping: example



Age	Salary
23	800
20	810
18	490
28	780
25	987
27	600

Age
18
20
23
25
27
28

Age
20
18
27
28
23
25

Age	Salary
27	800
18	810
20	490
25	780
28	987
23	600

Microaggregation

- Build small microclusters and replace each original data by its cluster representative.
- Usual method:
  - Partition: group similar records into cluster of at least k records
  - Aggregation: for each compute the cluster representative (centroid)
  - Replacement: each record is replaced by the cluster representative.

33

Micoraggregation: example

Partition k=2

Aggregation, replacement

Age	Salary
23	800
20	810
18	490
28	780
25	987
27	600

Age
23
20
18
28
25
27

Age
24
19
19
27.5
24
27.5

Age	Salary
24	800
19	810
19	490
27.5	780
24	987
27.5	600

34

Microaggregation: some notes

- Can be applied to multiple variables at the same time (multivariate).
- Can be used to achieve *k*-anonymity

35

Microaggregation: example (multivariate)

Weight	Age	Pathogen
70	33	Virus
84	40	Bacteria
59	15	Virus
69	35	Protozoa
63	25	Virus
60	20	Bacteria

Microaggregation
k=3
applied to quasi-identifiers: Weight and Age

Weight	Age	Pathogen
74.3	36	Virus
74.3	36	Bacteria
60.6	20	Virus
74.3	36	Protozoa
60.6	20	Virus
60.6	20	Bacteria

3-anonymous table

36

Additive and Multiplicative Noise

- **Additive Noise**: add noise to the original data:
  X′ = X + ε
  - ε is noise, following a certain distribution
- **Multiplicative Noise**: multiply noise to the original data:
  X′ = X * ε
  - ε is noise, following a certain distribution

37

Additive Noise: example

Weight	Age	Pathogen
70	33	Virus
84	40	Bacteria
69	35	Protozoa
59	15	Virus
63	25	Virus
60	20	Bacteria

$$X' = X + N(0, p\,\sigma^2(X))$$

Noise as a Normal Distribution with mean $\mu = 0$, standard deviation $\sigma^2 = p\,\sigma^2(X)$, with parameter *p=0.2*
Applied to Weight and Age

Weight	Age	Pathogen
71	36	Virus
81	40	Bacteria
68	30	Protozoa
58	16	Virus
62	29	Virus
59	23	Bacteria

38

Additive noise: example comments

- Bigger *p*, means more distortion, more privacy.
- This example uses **uncorrelated** noise:
  - Preserves means an covariances from the original data
- Multiplicative noise with a Laplace distribution is commonly used to provide differential-privacy.

39

Generalization

- Generalization of values following a given scheme or hierarchy.

Age	ZIP	Profession
30	08193	Nurse
36	08176	Paramedic
32	08191	Veterinarian
25	08034	Mathematician
29	08022	Physics
27	08010	Engineer

Age	ZIP	Profession
[30, 40)	081**	Health prof.
[30, 40)	081**	Health prof.
[30, 40)	081**	Health prof.
[20, 30)	080**	Science prof.
[20, 30)	080**	Science prof.
[20, 30)	080**	Science prof.

40

Evaluation of protection methods

- **Privacy**: level of privacy or anonymity achieved by the protection.
  - Privacy model: does it guarantees some privacy model? (k-anonymity, differential-privacy)
  - Reidentification: simulate an attack. How many users can the attacker re-identify?

Usually greater privacy implies greater distortion or generalization of the original data.

- We need to measure how data is deteriorated

41

Utility and Information loss

Information loss (IL) determines the information lost due to the anonymization.

$$IL(X, X') = divergence(X, X')$$

divergence is a way to compare both datasets X, and X'

- If X is very similar to X': lower IL
- If X is very different from X': higher IL

42

Information loss: Example

- **Mean Square Error (MSE)**: [for numerical data] consider the dataset as a matrix M and the protected version as M', and c(M) is the number of elements in the matrix M, then:

$$\text{divergence1}(M, M') = MSE(M, M') = \frac{\sum_{ij}(M_{ij} - M'_{ij})^2}{c(M)}$$

43

Information Loss: example MSE

X = M		X' = M'	
30.0	2.0	29.0	2.5
22.0	2.0	23.5	1.5
25.0	1.0	23.5	1.5
28.0	3.0	29.0	2.5

$$MSE(M, M') = \frac{\sum_{ij}(M_{ij} - M'_{ij})^2}{c(M)}$$

Microaggregation k=2

$$= ((30.0 - 29.0)^2 + (2.0 - 2.5)^2 + (22.0 - 23.5)^2 + (2.0 - 1.5)^2 + (25.0 - 23.5)^2 + (1.0 - 1.5)^2 + (28.0 - 29.0)^2 + (3.0 - 2.5)^2) / 8 = \mathbf{0.9375}$$

44

Information Loss: example MSE 2

X = M		X' = M'	
30.0	2.0	26.5	2.0
22.0	2.0	26.5	2.0
25.0	1.0	26.5	2.0
28.0	3.0	26.5	2.0

$$MSE(M, M') = \frac{\sum_{ij}(M_{ij} - M'_{ij})^2}{c(M)}$$

Microaggregation k=4

$= ((30.0 - 29.0)^2 + (2.0 - 2.5)^2 + (22.0 - 23.5)^2 + (2.0 - 1.5)^2 + (25.0 - 23.5)^2 + (1.0 - 1.5)^2 + (28.0 - 29.0)^2 + (3.0 - 2.5)^2) / 8 =$ **4.875**

45

Information loss: example

- Other generic divergence measures can be used:
  - mean absolute error, mean relative error,
  - different between several static characteristics: between correlation matrices, etc.
  - They can be combined (aggregated)
- Specific measures for machine learning
  - Us a specific machine learning model (decision tree, clustering, regression, …) and compare the performance of the model using training data from the original data X and the protected data X'

46

Guillermo Navarro-Arribas

47